Bolt Beranek and Newman Inc.

AD A108324

LEVEL

Report No. 3486

The Assessment of Speech Quality

February 1977

Submitted to:
Defense Advanced Research Projects Agency

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER 3486	2. GOVT ACCESSION NO. AD-A108 324
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) THE ASSESSMENT OF SPEECH QUALITY	5. TYPE OF REPORT & PERIOD COVERED Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R. S. Nickerson, A. W. F. Huggins	8. CONTRACT OR GRANT NUMBER(s) MDA903-75-C-0180
9. PERFORMING ORGANIZATION NAME AND ADDRESS Bolt Beranek and Newman Inc. 50 Moulton Street Cambridge, Massachusetts 02138	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, VA 22209	12. REPORT DATE February 1977
	13. NUMBER OF PAGES 67
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited. It may be released to the Clearinghouse Department of Commerce for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Compression
Vocoders
Linear Predictive Vocoders
Speech Quality Evaluation
Multidimensional Scaling

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Various methods that have been used to assess the quality of speech are reviewed. These methods are organized under three general topics: unidimensional quality assessment, judging of individual speech qualities, and multidimensional scaling. The importance of effects attributable to speech material, talkers and listeners is emphasized. The desirability of objective measures of speech quality is noted and some

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. candidate measures are discussed. Finally, the relationship between quality assessment and intelligibility testing is briefly considered.

The Assessment of Speech Quality

R. S. Nickerson
A. W. F. Huggins

February 1977

This research was supported by the Defense Advanced Research Projects Agency under ARPA Contract No. MDA903-75-C-0180.

Submitted to:

Defense Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Att: Dr. Robert Kahn

THE ASSESSMENT OF SPEECH QUALITY

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	3
2.	WHAT IS SPEECH QUALITY?	8
3.	METHODS OF UNIDIMENSIONAL QUALITY ASSESSMENT	12
	3.1 Isopreference Method	12
	3.2 Relative Preference Method	23
	3.3 Absolute Preference or Rating-Scale Method	25
	3.4 Category-Judgment Method	25
	3.5 Forced-choice Similarity Judgment Method	27
4.	JUDGING INDIVIDUAL SPEECH QUALITIES	29
5.	MULTIDIMENSIONAL SCALING	34
6.	SPEECH MATERIAL AND TALKER EFFECTS	39
7.	LISTENER AND OTHER EFFECTS	48
8.	THE POSSIBILITY OF OBJECTIVE MEASURES OF SPEECH QUALITY	50
9.	QUALITY AND INTELLIGIBILITY	54
	9.1 Quantification of Intelligibility	58
10.	CONCLUDING COMMENT	61
	REFERENCES	63

LIST OF ILLUSTRATIONS

Figure 1 14

Figure 2 17

Figure 3 20

Figure 4 47

LIST OF TABLES

TABLE 1 31

Abstract

Various methods that have been used to assess the quality of speech are reviewed. These methods are organized under three general topics: unidimensional quality assessment, judging of individual speech qualities, and multidimensional scaling. The importance of effects attributable to speech material, talkers and listeners is emphasized. The desirability of objective measures of speech quality is noted and some candidate measures are discussed. Finally, the relationship between quality assessment and intelligibility testing is briefly considered.

I. INTRODUCTION

In evaluating any speech-production or speech-transmission system, the first question that must be considered is whether what is produced or transmitted is understandable to the listener. The primacy of this question is obvious; if the speech is unintelligible, anything else that can be said of it is of little consequence.

Unfortunately, intelligibility appears to be a necessary, but not a sufficient, condition for acceptability. Speech that is highly understandable may be objectionable to a listener because of qualitative properties that have little to do with its intelligibility. Moreover, even when the speech produced by different systems is not equally intelligible, it is not safe to assume that the more intelligible speech will invariably be what a listener prefers (Beasley, Zemlin & Silverman,1972; Zemlin, Daniloff & Shriner, 1968). A completely satisfactory explanation of why this is the case probably requires a deeper understanding of the psychology of language than we currently possess. That qualitative factors do play a role in determining how people react to speech is clear, however. A voice with a fundamental frequency of 300 Hz is likely to be reacted to quite differently if it is perceived to be coming from an adult male speaker than if it is believed to be coming from a female or a young child. Speech that is monotone, or otherwise lacking of normal rhythmic structure, may be particularly grating on a listener. Qualities that appear to be symptomatic of illness (head cold, laryngitis), congenital anomalies (cleft palate, congenital deafness), heightened emotions (anger, fear), or drug-induced states (intoxication, sedation) may arouse reactions that are independent of how intelligible the speech is, or of the message it conveys (Kramer, 1963).

While the distinction between intelligibility and speech quality is an important one, the line should not be drawn too sharply. Sometimes intelligibility and quality problems may have a common underlying cause, as when inappropriate control of the velum results in mispronunciation of stop or nasal consonants and also gives the speech an overall nasal characteristic. And quality may indirectly affect intelligibility because of the attitude it engenders in the listener; speech that sounds peculiar may not be understood because the listener becomes so preoccupied with the strangeness of the sound that he fails to listen to the words. But it seems clear that the concepts of intelligibility and quality do differ, and that both are relevant to the assessment of speech.

The advent of synthesized speech adds a new dimension to the problem of quality evaluation. In the past, a listener could always assume that the speech he heard had been produced by a human being. It may have been modified, and perhaps degraded--and in some cases made to sound nonhuman--by a transmission process, but that it originally was emitted by a human speaker was never in question. Now that machines are learning to talk, however haltingly, the listener may no longer be so certain that he is listening to a person rather than to a machine. Moreover, it cannot be assumed that machines will always sound like machines and people like people. And it seems quite possible that the reaction that speech evokes from a listener may depend, to some degree, on whether it is perceived as having been produced by a machine or by another human being. It is conceivable, for example, that the same acoustic signal may be reacted to differently if it is assumed to have been produced by a human speaker and transmitted over a poor communication system than if it is assumed to have been produced by a machine and transmitted over a high-fidelity system. Such a finding would be in keeping with the results of studies that have shown that how noisy a sound is perceived to be, or how annoying it is, may depend on what is assumed to be emitting it (Cederlof, Johnson, & Sorensen, 1963; Kerrick, Nagel, & Bennett, 1968; Robinson, Bowsher, & Copeland, 1963).

The "originator" problem is complicated by the fact that among the most promising techniques that are currently being developed in efforts to minimize the bandwidth requirements for transmitting speech are some that blur the distinction between human- and machine-generated speech. These techniques involve one or another variant of what is generally referred to as the analysis-synthesis, or vocoder, approach. For an introduction to this approach to speech transmission and other digital speech-processing techniques, the reader is referred to Schroeder (1966), and to Bayless, Campanella, & Goldberg (1973). A thorough treatment of speech analysis and synthesis has been presented by Flanagan (1972). The application of linear predictive coding (LPC) techniques to speech vocoding has been discussed in detail by Makhoul and Wolf (1972).

Briefly, the vocoder approach to speech transmission involves a trading of computation for transmission bandwidth. A key element of the approach is a model of the speech-production system, which, when given appropriate inputs (an excitation signal and a set of time-varying parameters) will emit speech. Computation is required in the analysis phase of the process, during which the speech signal is subjected to a variety of analyses in an attempt to determine what parameter values would have to be applied to the model in order to produce that particular signal. These parameter values are then transmitted to the receiving node of the communication link. There they are fed to a synthesizer, which embodies the model that is being used, and speech is produced. In general, the more sophisticated the speech-production model, the greater the amount of computation that is required to determine the necessary parameter values, but the fewer the number of bits that must be transmitted per unit time to produce speech of a given quality.

What is of interest about this approach, for the moment, is the fact that although the speech originates with a human speaker, what the listener hears has been produced by a machine. Moreover, from the listener's vantage point, the involvement of the human is not essential; the same speech could have been produced by feeding to the synthesizer the appropriate model parameters from any other source that was capable of generating them, such as, for example, a computer program. Thus, a listener cannot tell, simply by listening, whether the speech he hears originated with a human being or with a machine.

Our purpose in this paper is to review various methods that have been used to assess speech quality. It should be noted at the outset that all these methods (with one exception) are open to two criticisms. First, the usual purpose of quality tests is to permit an informed choice of one system of speech transmission, music reproduction, hearing aid, etc, over another. In bringing the test into the laboratory, the desire to predict the quality of the system <u>in use</u> has been dropped. Instead of using the systems, subjects <u>make judgements</u> about them. A judgment task may make quite different demands of a system than its intended use, and no studies have been reported justifying the extrapolation of results of judgment tasks to real life situations. Furthermore, the materials played through the systems for judgment (especially for speech transmission systems) tend to be formal readings of prepared texts - citation-form speech - in place of the careless and rapid speech typical of conversations. The second major problem is that choices made on the basis of the quality-judgment tests are virtually never validated by subsequent tests under operational conditions.

Measuring the quality of speech is much more subjective than measuring its intelligibility. Quality that is adequate for one purpose, such as receiving stock prices over the phone, may be quite inadequate for another, such as carrying on a lengthy conversation with a friend. As a result of these, and other, difficulties, the problem of quality assessment has received less attention than that of intelligibility testing, and consequently the techniques are less refined in the former case than in the latter. The work that has been done on quality evaluation has been motivated by various interests, among which are synthesized speech (Nye, Ingemann, & Donald 1975); vocoded speech (Huggins & Nickerson, 1975); speech heard through a reproduction system (Gabrielsson, Roserberg, & Sjogren, 1974), over a transmission system (McDermott, 1969), or through a hearing aid (Gabrielsson & Sjogren, 1974, 1975a, b); and the speech of deaf persons (Martony & Franzen, 1966).

2. WHAT IS SPEECH QUALITY?

Undoubtedly most people will agree that they can recognize qualitative differences in the speech of different people, or in speech transmitted through different systems. And they will be able to say, independently of its intelligibility, that one speech sample is "better" in some global sense than another. Nevertheless, in spite of the fact that the concept of speech quality is a meaningful one, it is not easily defined very precisely.

Operationally, quality has typically been assessed by means either of _preference_ judgments, or of judgments of _similarity_ to a standard. One might therefore define quality in these terms. But this does not entirely settle the matter, because each of these concepts has its own definitional problems.

Preference, for example, is an ambiguous concept. One must ask: preference for what purpose? And the criteria may be quite different when stating preferences for speech that is to be used:

- over the telephone in conversations with friends and relatives

(speaker identifiability may be an important factor in this case)

- on short recorded messages strictly for the purpose of conveying information
- for entertainment, e.g., recorded singing, reading of poetry, novels, etc.

It is interesting to note that investigators who have used preference judgments for speech evaluation have not always used the term the same way. Rothauser, Urbanek, and Pachl (1968), for example, tried to determine "which of two signals to be compared is preferred by an average listener as a source of information." Munson and Karlin (1962) asked listeners to choose which of two signals they would prefer to use for a telephone call.

Another problem associated with defining quality in terms of preference is the fact that preferences may change over time. What sounds strange or unusual on a first hearing may sound quite unremarkable after even a little exposure. Personal experience bears out this fact, and there is ample evidence that the affective properties of auditory stimuli in general (Heyduk, 1975) and of speech in particular (Pachl, Urbanek, & Rothauser, 1971) change with frequent hearing. In addition to having some interest from a theoretical point of view, such changes have obvious practical implications. For practical purposes, one wants to know not only how acceptable the speech from a given system is when one first encounters it, but also how one's perception of it may change with continued exposure to it.

It is often assumed that judgments of the similarity of pairs of stimuli are more stimulus determined, and less variable across subjects, than are preference judgments (Green & Rao, 1971; McDermott, 1969). The assumption is a plausible one because it is so easy to imagine situations in which one would expect to get a high degree of agreement among subjects on judgments of similarity but not on judgments of preference. Most people would probably find it easy to decide, for example, which of two circles of radius 5" and 10" is more similar to a third circle of radius 4"; but they would probably find it much more difficult to say which of these circles (5" or 10") they _preferred_. Of course, they might consider the choice to be difficult--or even silly--because it is of no consequence. Thus, in this case, the individual differences in the preference judgments might be attributed to the lack of any real preference for one circle over the other.

Thus it might appear that judgments of similarity to a standard would provide a better basis than judgments of preference for an operational definition of quality. However, there are two arguments against this position. First, not only does it presuppose an appropriate standard in terms of which to make the similarity judgment, but it also equates deviance from the standard with degradation in quality. This may be a reasonable assumption in the case of processed (e.g., vocoded) speech, because in this case one can use the unprocessed speech as the standard, and presumably any changes resulting from the processing would be for the worse. Consider, however, the problem of assessing the quality of the speech of a deaf child. The ideal standard, in this case, would be the unimpaired speech of the same child, but that is not available. And it is probably not safe to assume a monotonic relationship between the degree of similarity between the speech of person A and that of person B, and the quality of that of person A.

A second problem with using similarity judgments as the basis for defining quality is the possibility that in so doing, one may define away the very thing that is of greatest practical concern. It is not enough to know, in evaluating a speech sample, whether it sounds similar to another sample; one wants to know whether it sounds "good" in some global sense. That these are not the same things may be seen by returning to our circle-preference illustration, and substituting for the two test circles an orange and a banana, and for the reference circle a tangerine. Again, one would expect a high degree of agreement among people in judging the orange to be more similar than the banana to the tangerine. One might expect much less agreement, however, on the question of which is preferred, the orange or the banana, and in this case the preferences would probably be meaningful.

We do not pretend to solve the problem of defining quality in this paper. Perhaps it is not solvable, except in an arbitrary way. It does seem important, however, to be aware of the fuzziness of the concept and alert to the difficulties that one can encounter in trying to get a consistent view of work on quality evaluation if this is not borne in mind.

3. METHODS OF UNIDIMENSIONAL QUALITY ASSESSMENT

One concerted effort to develop guidelines for speech-quality evaluation has been made by the Methods of Subjective Measurement Subcommittee of the Audio and Electroacoustics Group Standards Committee of the IEEE. Following six years of working on the problem, the subcommittee published their findings and conclusions as the "IEEE Recommended Practice for Speech Quality Measurements" (IEEE, 1969). Although the subcommittee noted that speech can be appraised in terms of a variety of factors (e.g., preference, loudness, intelligibility, recognizability of properties of the speaker's voice), it limited its attention in the Recommended Practice to preference measurements only. Three methods for obtaining such measurements--the Isopreference Method, the Relative Preference Method, and the Category-Judgment Method--were discussed in some detail. The committee pointed out, however, that each of these methods has limitations, and concluded that a method has not yet been developed that is generally applicable.

3.1 Isopreference Method

The isopreference method of speech-quality evaluation was originally developed by Munson and Karlin (1962). It, or a variant of it, has subsequently been used by several investigators (Bricker, 1963; Rothauser, Urbanek, & Pachl, 1968; Tedford & Frazier, 1966). As introduced by Munson and Karlin, the method involves two conceptually distinct procedures: (1) the determination of "isopreference contours," and (2) the development of a scale in terms of which the relationships between contours can be represented. An isopreference contour is plotted on a two-dimensional graph, one axis of which represents speech level and the other noise level. A contour connects all points representing equally preferred combinations of speech and noise levels. Figure 1 illustrates a hypothetical set of such contours.

The procedure used by Munson and Karlin for mapping an isopreference contour involves an iterative pair-comparison task. To establish a new point on a contour, one uses as a reference a speech-noise combination that corresponds to a point already known to be on the contour (the initial point may be chosen arbitrarily) and tries to find another combination that is equally preferred (i.e., selected over the reference 50% of the time).

The search for the new combination is confined to a region of the speech-level noise-level space that is in the immediate vicinity of the reference stimulus. A speech level (or a noise level) is chosen that is different--but not greatly different--from that of the reference, and combined with several noise (or speech) levels to define a set of test stimuli. All of the test stimuli are then matched with the reference stimulus in a series of pair comparison trials. That test stimulus which is preferred to the reference stimulus by 50% of the subjects (it may be necessary to interpolate between a stimulus that is preferred more than 50% and one that is preferred less than 50% of the time) is taken as the next point on the isopreference contour.

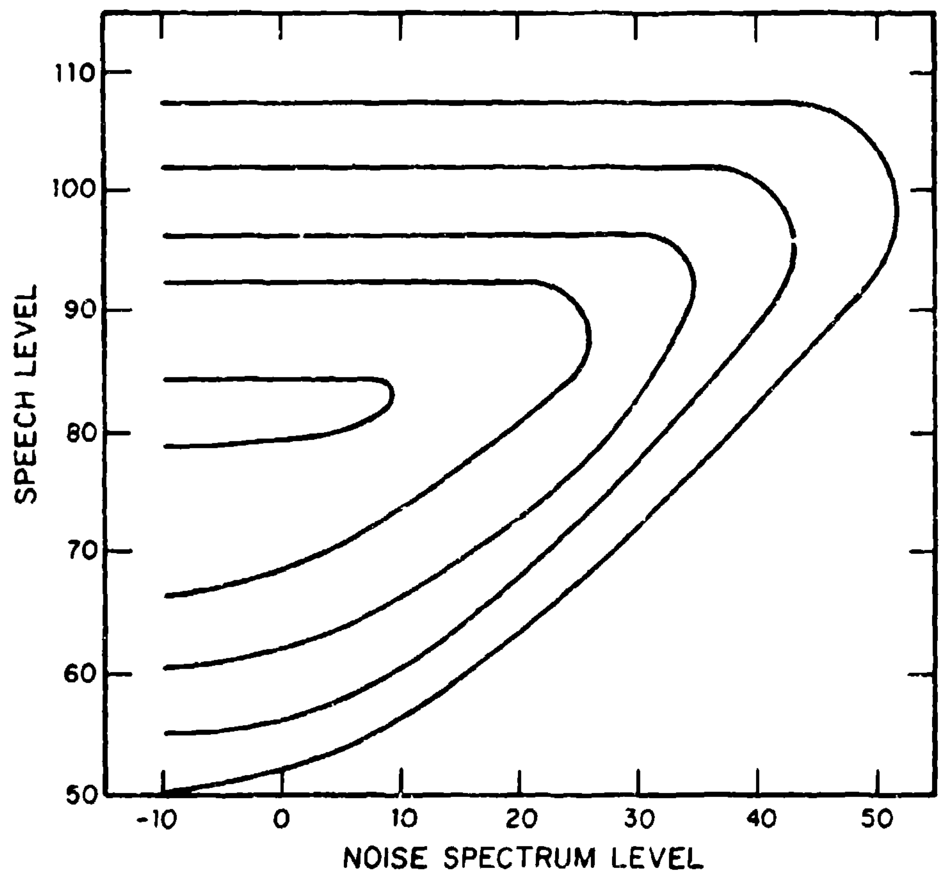

FIG. 1. A HYPOTHETICAL SET OF ISOPREFERENCE CONTOURS. ADAPTED FROM MUNSON AND KARLIN (1962).

(In Bricker's [1963] modification of this procedure, listeners are asked to produce equivalent impairments directly by adjusting noise levels to compensate for fixed differences in speech level.)

Several observations are worth making about isopreference contours, in addition to the fact that all speech-noise combinations represented by points on the same contour should be equally preferred. First, the smaller the area enclosed by a contour, the greater the preference for the speech-noise combinations represented by points on that contour. Or, to say the same thing in another way, all speech-noise combinations represented by points falling within an area enclosed by a contour should be preferred to all combinations represented by points falling outside that area. Second, the shapes of the contours indicate that for a given noise level, the speech level can be either too high or too low. Presumably, the criterion at the low-level end is influenced by the effect of the noise on intelligibility, whereas, at the high-level end it probably is less influenced by intelligibility and more by the annoyance of loud sounds. Note that the S/N ratio for two points on a contour with the same abscissa value may be very different. The fact that one can equate for preference speech that differs in such a striking way has been viewed as one of the main advantages of this approach. Third, the fact that the upper arms of the contours are flat suggests that when the speech level is sufficiently high, preference is relatively insensitive to noise level, provided the latter is moderate or low.

A fourth fact of some interest is that the ordinate value of an isopreference contour at the point at which the abscissa reaches its maximum value for that contour (the rightmost point of the contour) specifies the optimal speech level for a particular level of noise. It follows that a set of contours permits one to determine the optimal setting of speech level as a function of noise level. This relationship is given by a curve passing through the contours at their rightmost points, as illustrated in Fig. 2.

A set of contours, such as that illustrated in Figs. 1 and 2, shows how speech-level and noise level can be jointly varied within a group of equally preferred signals. It does not, however, provide any information concerning the relative preferences across groups, except their ordering. Addressing themselves to this problem, Munson and Karlin proposed two empirically-derived scales representing listener preferences numerically, a Transmission Preference Level Scale for which the scale values depend upon level of noise in a reference signal, and a Transmission Preference Unit Scale, which takes into account the variability of preference judgments. The difference between the ratings of two transmission systems, on the latter scale, may be used to predict the proportion of listeners that would prefer the system with the higher rating.

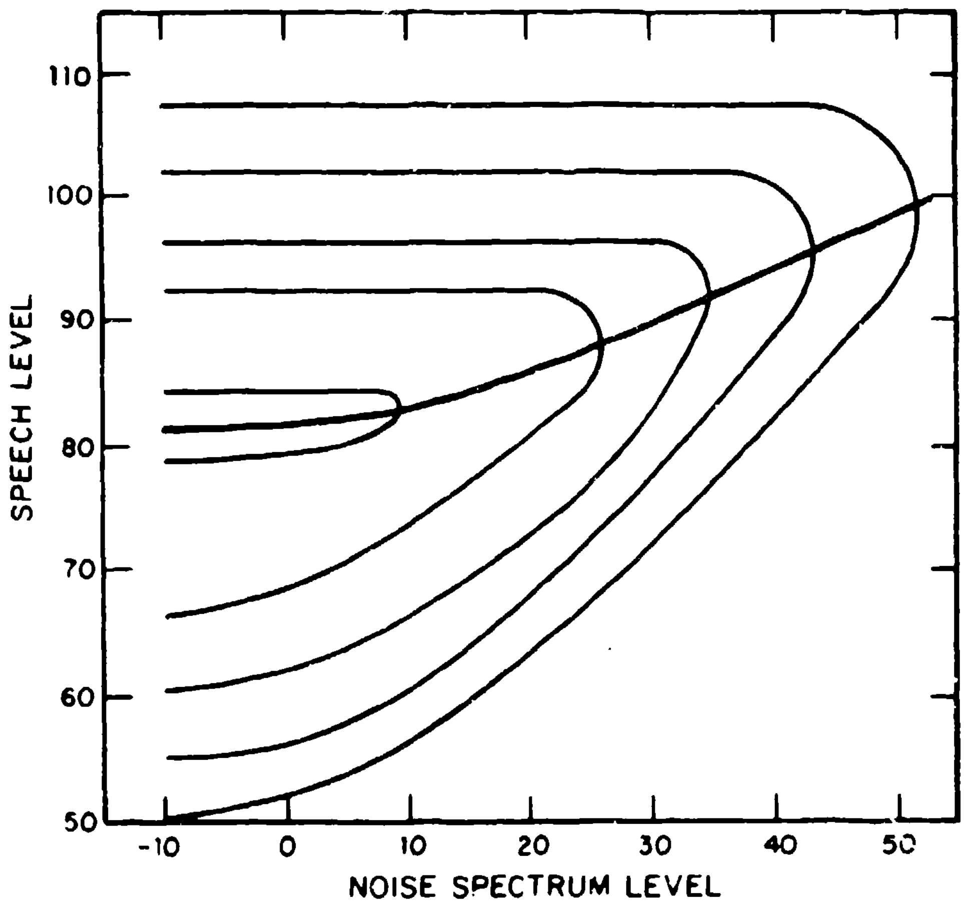

FIG. 2. A HYPOTHETICAL SET OF ISOPREFERENCE CONTOURS. THE SLOWLY ASCENDING CURVE REPRESENTS THE INFERRED OPTIMAL SPEECH LEVEL AS A FUNCTION OF NOISE LEVEL.

The isopreference method, as developed by Munson and Karlin has come under some criticism. Rothauser, Urbanek, and Pachl (1968) note that the method yields good results when the systems that are compared are represented by points that are relatively close on an isopreference contour; they point out, however, that deviations from predicted results become large when very high and very low speech-level systems on the same contour are compared. For this reason they recommend that the levels of both the reference and test speech signals be held relatively constant at empirically-determined optimum values, and that only the S/N ratio of the reference signal be varied during a single run of a quality-evaluation test.

Rothauser and his colleagues (Rothauser and Urbanek, 1965; Rothauser, Urbanek, and Pachl, 1968) have also argued against using additive white noise (as Munson and Karlin had done) as a means of degrading speech quality for preference testing. They argue that the signals that are produced by adding white noise to high-quality speech differ considerably from the output signals that are produced by most speech-processing systems. Moreover, they note, listeners may learn to separate such signals into their speech and noise components, an accomplishment that is facilitated by the fact that the noise is present during speech pauses. They advocate multiplying the noise source into the speech signal. (They also advocate the use of A-weighted pink noise, 3dB per octave attenuation of higher frequencies, in preference to white noise.) There are two advantages of the use of multiplicative noise over the use of additive noise: (1) the noise is present only when speech is present and is otherwise better integrated, perceptually, with the speech, and (2) computation of S/N ratio does not require measurement of the speech level in the former case as it does in the latter.

A third possibility, described both by Rothauser et al and by Schroeder (1968) involves adding to each digitized speech sample a noise derived directly from the speech by randomizing the sign of the sample. This additive noise has an intensity envelope that is identical with that of the original speech, with the result that the signal/noise ratio is the same for all speech sounds, and does not depend on measuring the speech level.

Rothauser, Urbanek and Pachl (1968) compared the effects of using additive and multiplicative noise directly. Figure 3 illustrates the type of results as an "isopreference curve" which is to be distinguished from Munson and Karlin's "isopreference contour." Each point on the curve gives the S/N ratio with additive noise that was equal in preference to a given S/N ratio with multiplicative noise. The results showed that over the range from -10 to +20 dB S/N, additive noise requires a higher signal to noise ratio (that is, additive noise must be relatively quieter) than multiplicative noise. Multiplicative noise has an advantage of about 3 dB for S/N ratios between about 0 and +10dB, increasing to 7dB or more as S/N ratio is either increased or decreased.

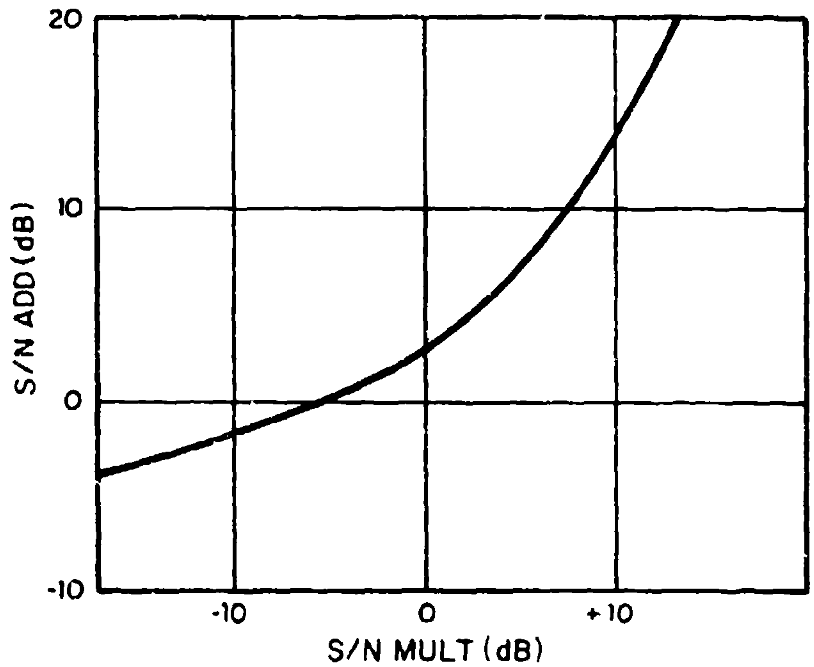

FIG. 3. AN EXAMPLE OF AN ISOPREFERENCE CURVE SHOWING THE RELATIVE EFFECTS OF ADDITIVE AND MULTIPLICATIVE NOISE ON JUDGED SPEECH QUALITY. ADAPTED FROM ROTHAUSER, URBANEK, AND PACHL (1968).

This result is not surprising, if the accepted method for measuring S/N ratios is considered. To achieve a 0dB S/N ratio with the additive noise, the noise is adjusted to have the same level as the speech <u>during the vowel peaks</u>. Since additive noise does not vary in level with the speech, this means that consonants are presented at S/N ratios of -10 to -30dB, depending on the particular consonant. With the multiplicative noise, on the other hand, the noise level is always correlated with the instantaneous speech level, and both vowels and consonants are presented at 0dB. The picture would be quite different if S/N ratios for both types of noise were defined relative to the average, rather than the peak levels.

In passing, a further criticism can be made of Rothauser et al's result. They claim to be measuring the relative preference for the two types of noise, as distinct from their effects on intelligibility. With S/N ratios below 0dB, however, intelligibility is bound to be impaired, and it is doubtful whether subject can ignore this fact while judging "quality."

An assumption underlying the isopreference approach is that overall quality differences can be represented adequately by differences in noise level or in S/N ratio. This assumption - whether the noise is additive or multiplicative, and however shaped - is difficult to accept. It is important to note, however, that the method does not require one to assume that qualitative differences could not be perceived between a reference signal degraded by noise and an equally-preferred test signal; it requires only the assumption that the two signals fall on the same point on a unidimensional preference scale. This assumption implies transitivity of the preference judgments: if listeners are indifferent to A and B and to B and C, they should also be indifferent to A and C when these are compared directly. The results obtained by Munson and Karlin were consistent with the transitivity requirement to within the error of measurement.

A genuine limitation of the method is the fact that it provides no clues concerning why any given signal is preferred to any other. If relative overall preference is the only question of interest, this limitation is of no consequence; if, however, one wants to know what must be done to a signal to improve its quality, then it is a major one.

Munson and Karlin warned against the possibility that the method could produce artifactual results, noting that "the experimenter must take precautions to ensure that the test does not revert to a simple discrimination problem in which the observer concentrates on detecting changes in the parameter of a transmission system instead of making a new and independent preference judgment on each pair of conditions" (p. 767). The problem stems from the fact that it is always clear to the listener which signal is the reference, and if he is motivated to be consistent in his responses, rather than to make an independent preference judgment on each trial, he may be able to do so. The proper precaution, according to these investigators, is to schedule the pairs that are to be compared in such a way that the listener is not likely to detect systematic changes in the reference or test signal.

3.2 Relative Preference Method

Hecker and Williams (1966) suggested, and tested, the possibility of using as reference signals, speech that had been distorted in fundamentally different ways, rather than the same speech degraded by different levels of noise. The method they proposed has come to be known as the "relative preference method." This method involves two steps: the development of a quality scale based on pairwise comparisons among a set of reference signals, all but one of which have been distorted in different ways, and the positioning of a test signal on that scale. Reference signals are selected so as to insure a range of quality. The types of distorting operations used by Hecker and Williams included bandpass filtering (300-3000 Hz), low pass filtering (3000 Hz) plus mixing with low-pass filtered (500 Hz) white noise, mixing with reverberant echo, and peak clipping (30 dB) combined with bandpass filtering (300-2000 Hz). Testing involves matching every reference signal (the IEEE subcommittee recommends the use of five of them) with every other signal as well as with the test item. The scale is constructed by ordering the reference items in terms of the relative frequency with which each is preferred to the others with which it was matched. The test item is then located on the resulting scale on the basis of the percentage of reference items over which it is preferred. Thus, a test signal that is preferred to 70% of the reference signals would be located above a reference signal on the scale that is preferred to 60% of the other reference signals, and below one that is preferred to 80% of them. Ideally reference signals should be equally spaced in terms of quality. At a minimum it must be the case that transitivity holds; otherwise the relationships among the reference signals could not be represented by a unidimensional scale.

The procedure does not guarantee the transitivity of preferences involving comparisons between test signals and particular reference signals. Inasmuch as the positioning of the test signal on the scale is determined solely by the percentage of reference signals to which it is preferred, the possibility that it may appear below a particular reference signal to which it is preferred, or above one that is preferred to it, is not precluded. Such an outcome would suggest, however, either that the scale is invalid, or that the judgments between test and reference items are being made on a basis that cannot be represented by a unidimensional scale.

Hecker and Williams compared the performance of listeners who gave preference judgments with the relative preference method against that of listeners who were tested with the isopreference technique. They found less interlistener variability in the former case, and concluded that that evaluation technique permitted more efficient preference testing that the conventional isopreference method.

3.3 Absolute Preference or Rating-Scale Method

This method requires that the listener assign to each test signal a number that represents his opinion concerning where that signal should be placed on a scale of speech quality or listener preference. The scale may be constrained to have a finite number of points--usually from five to ten--or the listener may be allowed to use fractional numbers as ratings and thereby make as finely graded distinctions as he wishes.

Pachl, Urbanek and Rothauser (1971) have shown that listeners may give different results in rating tests when they are explicitly instructed to use the highest rating with the best test items and the lowest rating with the worst, than when not given such instructions. When not instructed in this way, their listeners tended to make little if any use of the extreme values of a 5-point scale. This reluctance to use the upper extreme of a rating scale seems to be borne out by the results of a study by Gabrielsson and Sjogren (1975b) in which listeners were asked to rate speech heard through a hearing aid and speaker on a scale from 0 ("practically no fidelity at all") to 10 ("perfectly true to nature [sounds like the original sound]"). The mean rating for the minimal-distortion control condition (100-15000 Hz $\pm$ 3dB, < 1% distortion) was 7.7.

3.4 Category-Judgment Method

The listener's task in this case is to place each test signal in one of several categories representing specified levels of quality:

e.g., excellent, good, fair, poor, bad. Superficially, at least, the task is identical to that of the rating scale method except that the ratings are represented by descriptive terms rather than by numbers.

As Grether and Stroh (1972) have pointed out, the Isopreference and Relative Preference methods involve the assumption that the listener's subjective assessment of speech quality can be appropriately represented by a unidimensional continuum. These investigators have argued that one should try either to establish what the relevant psychological dimensions of quality are, or to present evidence that use of a single composite dimension is reasonable. They suggest that in the absence of independent measures of the performance of a communication system against which to validate quality measures, quality assessment techniques should be judged in terms of three criteria: (1) simplicity (of the information processing demands placed on the listener), (2) relevance (degree of correspondence of the laboratory task to "real world" speech perception), and (3) reliability (repeatability of measurements obtained). Grether and Stroh contend that the Category Judgment Method satisfies these criteria. They recommend use of a 9-point scale with alternating points labelled Excellent, Good, Fair, Poor and Unsatisfactory and the remaining points unlabelled.

A practical difficulty with the method is that of assuring that all listeners interpret the category labels in the same way. The approach that typically is taken to attempt to come to grips with this problem is to try to anchor the scale on one or both ends--and possibly at other points as well--by presenting examples of what should be considered good signals, poor signals, or whatever. To control for the possibility that an individual's categorization criteria might wander as a result of exposure to many signals during the test, anchoring signals, along with their appropriate categorizations, may be presented several times while the test is being conducted.

3.5 Forced-choice Similarity Judgment Method

A still different approach to quality evaluation was tried by Mostofsky (1969). The listener's task in this case was to decide which of two "anchor" stimuli a test stimulus most closely resembled. The test stimuli were sentences that had been processed by a channel vocoder. The independent variable of interest was the quantization step for each vocoder channel. Nine levels of this variable were used, the smallest and largest step sizes being associated with transmission rates of 3200 and 1400 bits per second, respectively. The anchor stimuli that were used to mark the ends of the quality continuum were: (a) a sample of unprocessed speech, and (b) a sample of speech processed by the 1400 bits per second system.

On each trial, the subject was permitted to listen to either or both anchors as many times as he wished before making his decision. When he felt ready to do so, he then simply indicated which of the anchors was most similar in quality to the test stimulus.

Three performance measures were taken, the first two of which were considered to be indications of a subject's confidence in his judgment.

a. The number of times the subject listened to the anchor stimuli

b. Decision time

c. The anchor selected as most similar to the test stimulus

The results appeared to be relatively insensitive to differences in vocoder bandwidth; neither the number of references to the anchor stimuli nor decision time seemed to depend very much on this variable. The similarity judgments were divided into two groups. Unprocessed speech was judged to be similar to the unprocessed anchor. All vocoded samples, irrespective of the quantization step size, were judged to be more similar to the poor quality anchor than to the unprocessed anchor. In other words, all samples except those of the highest quality were assigned to the "poor quality" category. (This was true whether the samples were played in the normal, or reversed, direction.)

The most straightforward interpretation of the latter result is that the perceptual difference between the unprocessed speech and the best of the processed samples was larger than the difference between the best and worst samples of processed speech. A different result might have been obtained had the processed samples included some of higher bandwidth than 3200 bps. A fairly clear implication of the result, vis-a-vis the question of evaluation methodology, is that the technique appears not to be very sensitive to quality differences, given that the best of the processed speech samples differs appreciably from the unprocessed standard.

4. JUDGING INDIVIDUAL SPEECH QUALITIES

An approach to speech evaluation quite different from that of obtaining judgments of its overall quality is that of trying to assess it with respect to specific aspects or features. One may ask a listener to attend to one or more characteristics of an utterance, such as its loudness (Coolidge & Reir, 1959), its degree of nasality (Stevens, Nickerson, Rollins, & Boothroyd, 1974), the appropriateness of its timing or rhythm (Boothroyd, Nickerson, & Stevens, 1974), its pitch and intonation (Stratton, 1973), the degree to which it preserves the voice characteristics of the talker (Becker & Kryter, 1975).

A commonly used method for obtaining descriptions of complex stimuli in terms of several unidimensional properties is that of semantic differential scaling (Osgood, 1952). The method involves the rating of the same stimulus on several scales, each of which is defined in terms of a pair of antonymous words that designate its end points. One result that is obtained from this technique is a semantic differential profile which represents a description of a stimulus in terms of the dimensions of the analysis.

The approach is illustrated by an experiment by Kerrick, Nagel, and Bennett (1968), one objective of which was to determine the extent to which the concepts of loudness and noisiness could be operationally distinguished. Table 1 shows the scaling dimensions that were used in this case. Loudness and noisiness proved to be nearly equivalent descriptors in this study, the correlation between ratings on these dimensions being .96. A plot of the stimuli in a space, the coordinates of which were the noisiness and acceptableness continua, suggested that the acceptability of a given level of perceived noisiness depends on the nature of the sound; higher levels of noisiness were acceptable for musical sounds than for vehicle sounds, and for vehicle sounds than for "artificial" sounds.

An incidental, but suggestive, result from this study came from a comparison of the reactions of two listeners to the same sound (broad-band noise). Subjects were not told the source of the sounds but were asked to identify them. One subject identified this sound as "air blowing" and another as a jet flyover. The former subject judged the sound to be louder and noisier, but more acceptable, than did the latter, suggesting that the degree of acceptability of a given level of perceived noisiness may depend not only on the nature of the sound but also on that of its assumed origin. While this result was obtained with non-speech stimuli, it points out the importance of variables other than stimulus properties per se as determinants of individual preferences, and it seems likely that similar effects might be found with speech.

Table 1. Scaling dimensions used by Kerrick, Nagel, and Bennett (1968) for semantic-differential description of sounds. Listeners rated each sound with respect to each of these dimensions on a 7-point scale.

good	---	bad
far	---	near
unfamiliar	---	familiar
noisy	---	quiet
fast	---	slow
smooth	---	rough
natural	---	unnatural
soft	---	loud
passive	---	active
acceptable	---	unacceptable
high	---	low
delicate	---	rugged
pleasant	---	unpleasant
narrow	---	wide
light	---	heavy

Another example of the use of judgments with respect to specific properties of sounds comes from Gabrielsson and Sjogren (1974,1975). Their listeners rated auditory stimuli (including speech, but also symphonic music, household sounds and traffic noise) with respect to several (62 in one experiment, 40 in the other) "adjective scales." Examples of the adjectives that were used are distant, pleasant, brilliant, stark, dull. The listener's task was to rate each sound with respect to each adjective using a 10-point scale (0 through 9) to indicate the degree to which that sound had the quality designated by that adjective. The sounds that were judged had been passed through one of several hearing aids that were being evaluated.

Nakatani and Dukes (1973), in the course of testing their Q-measure described in more detail below, had subjects rate several properties of speech that had been passed through a variety of distortions, including two levels each of high- and low-pass filtering, and of additive noise, and telephone speech. The rated properties were distortion, noise, understandability, pleasantness, quality, and fidelity. They found that ratings on all of these scales <u>except</u> <u>noise</u> were highly intercorrelated (negatively in the case of distortion). The fact that noise was <u>not</u> highly correlated casts doubt on the fundamental premise on which Munson and Karlin's (1962) isopreference test is based: that a unidimensional comparison can be made between a speech sample of arbitrary quality, and a reference signal degraded by noise. A similar conclusion was reached by McDermott (1969: see below).

The approach of comparing speech with respect to specific characteristics has been criticized on the grounds that how to derive a measure of overall quality from the results of such comparisons is not known (Rothauser, Urbanek, & Pachl, 1968; Tedford & Frazier, 1966). To the extent that one is interested in differences with respect to specific features per se, as opposed to differences in overall quality, this limitation is irrelevant. But if the primary interest is in overall quality differences, it clearly is relevant.

Another problem with the approach is the paucity of evidence that people can make reliable judgments about a specific feature of an utterance, independently of its other features. Tedford and Frazier (1966) see the fact that the isopreference method does not require the listener to analyze his reasons for preferring one speech sample over another to be one of the major advantages of that approach.

Gabrielsson and Sjogren (1975a) note also the difficulty that some of their subjects nad in making their ratings of sound reproduction with respect to specific characteristics independently from the characteristics of the sounds per se. A variety of nonspeech sounds, in addition to speech, was used in this experiment, and it is hard to imagine that one could judge, say, the "shrillness", or the "dullness" of the reproduction of a sound without being influenced by the shrillness or dullness of the sound itself.

Still another problem that has been pointed out by Rothauser, Urbanek, and Pachl (1968) is that a given qualitative descriptor can mean different things to different listeners or in different contexts. When used in connection with synthetic speech, "naturalness," for example, might represent the degree to which the speech sounds human; whereas in the context of judging telephone circuits, the same term might be used to indicate the degree to which a transmission preserves the voice characteristics of a particular speaker.

In spite of these limitations, the comparison of speech samples with respect to specific characteristics can be a useful thing to do. It can be a particularly helpful approach when there is reason to believe that the difference between the overall quality of two systems is attributable to specific identifiable characteristics. And as was noted above, the identification of specific qualitative aspects of a system's output may sometimes be more useful to the developer of the system than non-specific information concerning global quality.

5. MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) methods attempt to model data by representing each stimulus, or vocoder system, as a point in an n-dimensional space, such that the data reconstructed from the model match the empirical data as closely as possible. There are several classes of models, which Carroll (1972) has shown to be hierarchically related, in that each class is a special case of the next-higher class in the hierarchy. The simplest is the vector model. Here, the data are represented by the ordering and relative spacing of the stimulus-points as they project onto a vector through the space. Each subject, or each condition under which data were collected, is represented by a different vector. A second class of models (the unfolding models) represents both stimuli _and_ subjects as points, and the subject's preferences are represented by the distances from "his" point to the various system points, the closest being the most preferred.

By doing a multidimensional analysis for different values of _n_, one can determine how many dimensions are necessary to account for the results at any given level of precision. Precision always increases, or it least does not decrease, as _n_ is increased. It is also the case, as McDermott (1969) points out, that reliability tends to decrease as dimensionality is increased, particularly when the dimensionality of the solution is greater than that of the stimuli, so the higher dimensions are accounting only for noise in the data.

Several points are worth emphasizing with respect to the solution space generated by an MDS analysis. First, the space that is used to model the data is a perceptual, or subjective one. Second, the analysis itself does not identify what the factors are that are represented by the coordinate axes of the space; it only indicates how well _n_ of them can account for the data. One can sometimes make a reasonable guess concerning what one or more of the axes represent by simply noting the way the stimuli are distributed throughout the space, but this is not always possible. Third, the subjective factors represented by the axes may or may not have physical correlates; that is to say it may or may not be possible to associate the axes of the subjective space with objective properties of the stimuli.

Both judgments of similarity and judgments of preference have been used as input data for MDS procedures. In keeping with the assumption that preference judgments are less stimulus determined, and more affected by individual differences, than are similarity judgments, scaling procedures applied to the former usually represent intersubject differences explicitly in the results of the analysis, whereas many of the procedures applied to the latter do not.

Only a few efforts have been made to apply MDS to speech evaluation. One such effort was a study by McDermott (1969), in which some listeners made pairwise similarity judgments (expressed on a 10-point scale), and others stated a preference for one item of each of the possible stimulus pairs. Stimuli were sentences processed through 22 different circuits. The tested circuits included a peak clipper, a center clipper, a full-wave rectifier, a chopper, an E. B. Bank (a very sharp low-pass filter), a frequency shifter, a vocoder, an echo, a comb filter, several noise and signal intensity levels, and several band-pass filters. Both types of listener judgments were subjected to MDS analyses. The distribution of systems in 3-dimension solution spaces were very similar for the two types of judgments. This finding suggests that both judgments were based on the same underlying stimulus features.

Positioning of the systems in the solution spaces suggested to McDermott that the 3 coordinates represented (1) overall speech clarity, (2) a dimension associated with whether circuit degradation resulted from signal distortion or background interference, and (3) subjective loudness. The positioning of the subject vectors in the preference space suggested that individual listeners differentially weighted different attributes in arriving at their preferences. The results suggested that most listeners tended to give greatest weight to overall clarity as the most preferred attribute, but that they differed considerably with respect to their weighting of the two types of degradation (signal distortion and background noise) that were used. McDermott concluded from this result that quality assessment techniques that average preference judgments over individuals have limited validity. In particular, she noted the limitations of methods that make use of the concept of equivalent single-parameter degradation (e.g. isopreference methods) to represent speech quality. "Although these methods have the important advantage of expressing quality as a single number on a unidimensional scale, the evidence from the present experiment suggests that these equivalent degradation methods can be subject to all the disadvantages of large amounts of inherent intersubject variability" (p. 781). She further concluded that to the extent that a unidimensional measure of equivalent quality is desirable, such a measure should correspond maximally with signal clarity and minimally with signal-noise distortions and loudness.

Other attempts to apply multidimensional methods to the analysis of speech-quality judgments have typically found no more than two, or three, and sometimes only one, perceptual dimensions underlying quality (Gabrielsson & Sjogren, 1975, McGee, 1964, 1965). In one study in which semantic differential data (15 scales) were factor analyzed, McGee (1964) found two roots to be significant, and he identified the corresponding factors as Intelligibility and Naturalness. In a second study (McGee, 1965) he found that a single factor accounted for most of the variance. Gabrielsson and Sjogren (1975), required three dimensions, however, to account for from 66% to 72% of the variance in similarity judgments made on symphonic music and speech that had been passed through one of several hearing aids and a loudspeaker. One of these dimensions was identified as a composite of brightness-darkness, fullness, loudness and perceived distance. A second dimension was identified as clearness or distinctness. The third was not given a perpetual label. An attempt was made to relate the perceptual dimensions to physical characteristics of the aids such as bandwidth, region of maximum response, locations and relative magnitudes of resonant peaks. The locations of the different aids in the perceptual space was not quite the same for speech material as for music.

Gabrielsson and Sjogren got only partial agreement between the results of the MDS analysis based on similarity judgments and the factor analyses based on ratings with respect to specific characteristics (see section 4). They point out that the experimenter may, in effect, determine the dimensions of the perceptual space in the latter type of experiment by selecting the descriptive adjectives in terms of which the subjects must respond; whereas this is not the case when similarity judgments are used. This type of finding demonstrates the need for more direct comparisons among different assessment methods using the same stimulus materials.

6. SPEECH MATERIAL AND TALKER EFFECTS

Relatively few studies have focused on the role of the nature of the speech material or the characteristics of the talker's speech as determinants of the outcomes of quality evaluations. Those that have, however, have shown that these effects can be substantial, and if not taken into account, can lead to faulty interpretations of results.

House, Williams, Hecker and Kryter (1965), for example, found a quite large talker effect in a study designed to assess the effectiveness of an intelligibility testing procedure. The difference in intelligibility of the words produced by the two talkers who recorded the test material was comparable to that resulting from a difference of 3 dB in signal-to-noise ratio. This difference may have been due to a difference between the two speakers in the relative levels of vowels and consonants (Horii, House, and Hughes, 1971). The signal/noise levels were determined relative to vowel levels, but the test was of consonant identification.

Voiers (1972) found a relationship between the intelligibility (as measured by the Diagnostic Rhyme Test) of vocoded speech and the fundamental frequency of the speakers voice, the higher intelligibility scores being associated with the lower fundamental frequencies. (Unfortunately only male speakers were used in this study and fundamental frequencies are not reported.) While this study concerned intelligibility rather than judged quality, it seems likely that had quality judgments been made they would have shown a similar effect. Voiers has concluded that digital vocoders, vintage 1972, affect speech perception in much the same way as does band-limited Gaussian noise. He notes that the performance of these vocoders tends to differ in systematic ways for voiced and unvoiced sounds; in particular <u>manner</u> of articulation is better preserved in unvoiced sounds, and <u>place</u> of articulation in voiced sounds.

Hirsh, Reynolds and Joseph (1954) got significant material and talker effects in a study of the intelligibility of masked or filtered speech. An interesting aspect of the results obtained with filtered speech was a talker-by-degree of distortion interaction that was attributable in part to the fact that words spoken by females were more intelligible than those spoken by males when the speech was high-pass filtered with a cutoff at 3200 Hz or above.

Evidence of the importance of proper selection of test sentences has also been presented by Pachl, Urbanek, and Rothauser (1971). In their study, the percentage of judgments favoring a given system over others in a direct comparison task varied greatly depending on the sentence that was used for the comparison. Pachl, Urbanek and Rothauser concluded from their finding that if meaningful results are to be obtained from preference judgments, the same test materials must be used with all systems. We agree with this point, but suggest that invariance of materials across systems is, by itself, an insufficient requirement. It is also essential that the material that is used with a given system be as broadly representative of the vagaries of speech as is practically feasible, and that the same broad sampling of material be used with every system. Use of material that is invariant across systems, but not broadly representative of speech in general, could yield misleading results by producing a rank ordering of systems that would hold only for speech with the particular characteristics of the sample used.

Our own work on quality evaluation began with the observation that one of the main causes of variability in quality testing is the difficulty of the subject's task. Judgements of global quality are not easy when the stimuli being compared differ in a variety of ways. Nor is it a simple matter to compare speech samples with respect to some particular property when they differ with respect to many other properties as well. One way to simplify the subject's task would be to arrange that the stimuli presented for judgement differ with respect to only one perceptual dimension at a time.

Note that this is not the same as asking the subject to abstract one dimension perceptually in order to compare stimuli with respect to that dimension when they differ in many other ways as well. We attempted to achieve this by analyzing the sources of the errors that the vocoding process introduces into speech, and targetting each of these sources with a sentence designed to maximize the errors due to it, while minimizing the errors due to the other sources. Thus, in contrast to earlier material, which aimed at phonetic balance, our sentences are Phoneme-Specific, in that they concentrate phonemes with similar acoustic properties in a single sentence.

Although our tests were aimed specifically at Linear Predictive (LPC) vocoders, the procedures that were developed are probably equally applicable to other methods of vocoding. An LPC vocoder first models the spectrum of a short sample of the waveform by calculating the parameters of an all-pole filter with the same spectrum. This introduces the first source of error: some speech sounds (e.g. nasals and fricatives) contain zeroes as well as poles in their spectra, and these may not be adequately matched by an all-pole model. Next, the coefficients that define the modelling filter are quantized. The quantization introduces a second type of error, which would be most likely to have an effect on perceived quality when the quantization steps are slowly swept, as in vowels and semi-vowels. Thirdly, the window defining the waveform sample is moved down the waveform by a time called the 'frame size' and the spectral modelling is repeated. The larger the frame size, the wider the intervals at which the speech spectrum is sampled, and the greater the chance that rapidly changing parts of the waveform will be represented inadequately. This type of error should be most noticeable in speech sounds that show rapid spectral and amplitude changes, such as the stops and affricates.

In view of these considerations, we composed a set of four Phoneme-Specific sentences, plus two "general" sentences that contained difficult clusters and strings of unstressed syllables. The sentences were as follows:

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea-party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

The six sentences were read by twenty talkers. From these, three males and three females were selected so as to represent the range of fundamental frequency, and degree of nasality, found in the whole group of twenty. The set of thirty six speaker-sentence combinations were then processed through a set of twelve LPC vocoder systems, whose number of poles, quantization step size, and frame size, were traded off against each other to equate the bit-rates of all systems to 2600 b.p.s. The 432 resulting stimulus sentences, together with a PCM version of each sentence, and a vocoded but unquantized version, were presented to well trained subjects in two separate tasks. In one, subjects rank ordered the systems separately for each of the 36 speaker-sentence combinations. The sentences were transferred to Language Master cards for this purpose. In the second task, all 504 stimulus sentences were presented in a counterbalanced order, and subjects rated the 'degradedness' of the speech, assigning larger numbers to more degraded systems.

Multi-dimensional scaling of the data, using MDPREF (Carroll, 1972), showed that different perceptual effects were associated with inadequate static spectral match and with inadequate dynamic spectral match. An inadequate static spectral match, resulting from too few poles, or from too coarse quantization, produced quality that could be described as "muffled", or as "burbly", respectively. Separation of the vocoder systems along these dimensions was achieved as a result of using speakers with a wide range of fundamental frequencies. On the other hand, an inadequate dynamic match, resulting from too long a frame size, produced a "chirpy" or "bleaty" quality, and separation along this dimension was the result of our choice of sentence materials. Furthermore, the foregoing two perceptual dimensions were orthogonal, suggesting that they were independent.

A further result was that the data from the ranking and rating tasks yielded highly similar solution spaces in the MDPREF analysis. This implies that the ranking and rating tasks are alternative and equivalent methods of measuring a single underlying perceptual sensory continuum, or set of continua. If so, it is probably appropriate to discard the less efficient procedure, in this case the ranking task.

Our results tend to corroborate those of other studies that have indicated the importance both of the words and sentences that are selected for testing, and of the voices that are used to record the test materials. The possible magnitude of sentence and talker effects is illustrated by Fig. 4, which shows mean preference ratings (4 listeners) over the same 14 LPC systems with two different talker-sentence combinations. Given the occurrence of such effects, the need is apparent to use a broad range of sentence and talker characteristics in any test that is intended to compare systems with respect to overall quality. The possibility of the biasing of results due to inadequate sampling is quite real.

A second reason for using carefully selected materials representing a broad range of characteristics is the fact that doing so provides an opportunity for acquiring information about the strengths and weaknesses of individual systems to deal with specific aspects of speech or voice characteristics. This point also is illustrated by Fig. 4. Consider, for example, system 6. This system ranked close to best in preference with talker RS and sentence 1, but worst with talker AR and sentence 4. It clearly would be of interest to a system designer to know the cause of this difference. A consideration of the parameters of the system itself and of the characteristics of the talker and sentences provides some hints. The system was an LPC system with 10 poles (no zeros), a 25 msec frame size, a quantization step of 0.2 dB, and a constant transmission rate of 2650 bits per sec. Speech sample 1 ("Why were you away a year, Roy?") is voiced throughout and contains only vowels and /w, r, y/. These sounds are all characterized by slow rates of change of both spectrum and envelope. In short, this sentence is relatively "smooth" and free of abrupt changes. In contrast, sentence 4 ("Which tea-party did Baker go to?") contains many stop consonants and affricates, which are characterized by very abrupt changes in both spectrum and envelope. Talker RS is a female with a moderate speaking rate and an average (209 Hz) fundamental frequency, whereas talker AR, also female, talks rapidly and (for a female) has a relatively low (167 Hz) fundamental frequency. In the light of these facts, the results shown in Fig. 4, as they pertain to system 6, are not so surprising. The system was apparently able to give adequate coding of a slowly changing spectrum, as in RS-1, but was unable to cope with the repeated abrupt changes in AR-4.

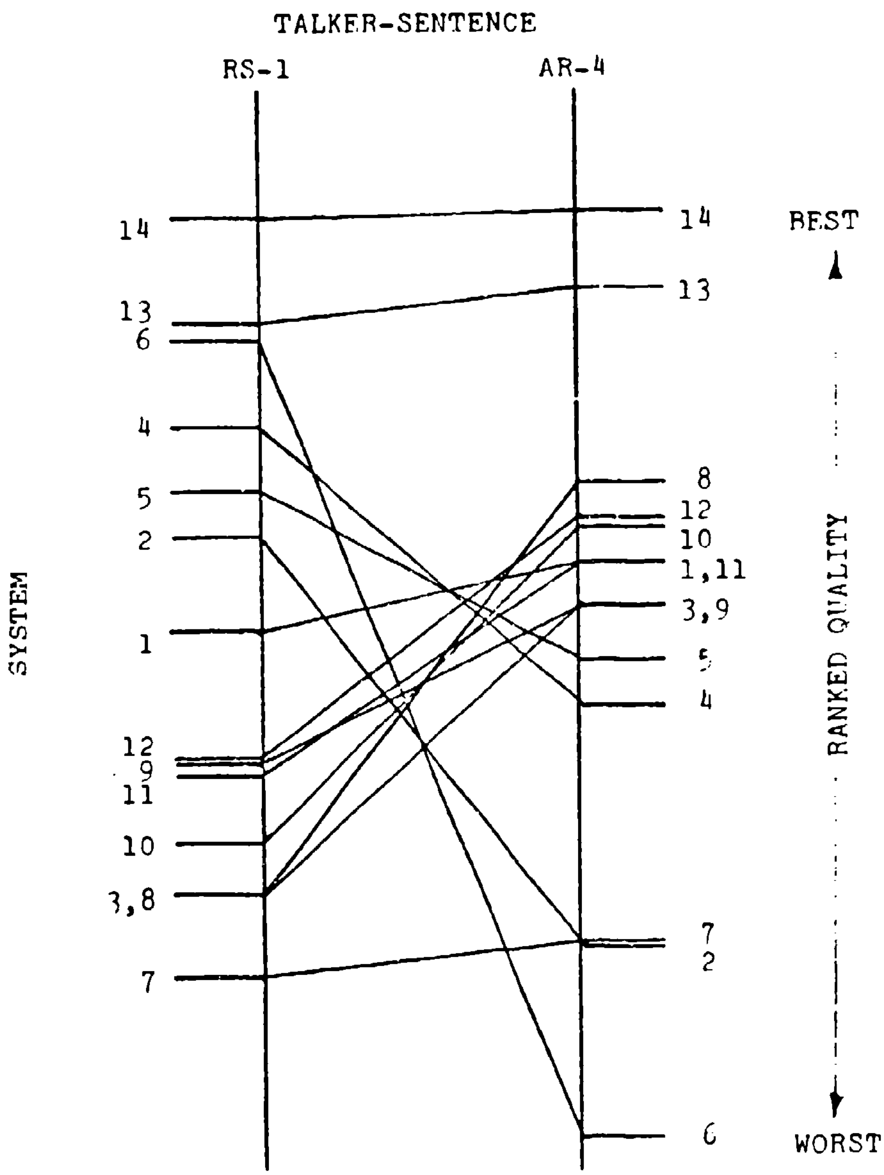

FIG. 4. MEAN QUALITY RATINGS OF SPEECH PROCESSED BY 14 LPC SYSTEMS WITH TWO DIFFERENT TALKER-SENTENCE COMBINATIONS. SEE TEXT FOR EXPLANATION.

It seems clear from these results that if one's purpose is to determine the quality of the output of a given system, or to compare several systems, the use of a wide sampling of both speech material and speech characteristics is imperative. If the purpose is to develop or test an evaluation procedure, as, for example, in the Munson and Karlin (1962) study, one may get by with a more restricted sample. As an aside we might note also that a system that is judged to produce high quality non-speech material (e.g. music or environmental sounds) may not necessarily be judged to produce high quality speech (Gabrielsson & Sjogren, 1974, 1975).

7. LISTENER AND OTHER EFFECTS

A very important factor in determining either the intelligibility or the quality of speech is the degree to which the listener is familiar with the characteristics of the speech to which he is listening. Hudgins (1943, 1949) has exphasized this point in connection with the problem of assessing the intelligibility of the speech of deaf children. Listening performance in this case is sensitive to the listener's familiarity with (a) the speech of deaf persons, (b) the speech of the particular speaker, and (c) the material from which the speech samples have been drawn. Concerning the latter factor: it is clear that familiarity with specific test utterances improves listening performance; it seems highly probable that familiarity with the linguistic structure of the material would do so also. If, for example, the listener knows that test sentences are invariably active voice and have the structure noun phrase -

verb phrase - noun phrase, this knowledge should be helpful.

The existence of listener effects such as those noted above are particularly relevant to the problem of evaluating the performance of speech-vocoding systems. One implication is that the worst possible judges of the intelligibility or quality of the speech that any given system is producing are individuals who are intimately involved with the development of that system, and consequently familiar with the characteristics of its speech. On the other hand, it is not necessarily the case that as judges one wants listeners who are representative of the population in general. Rothauser, Urbanek and Pachl (1968) contend that the most appropriate judges of the quality of a speech transmission system are listeners who are representative of the intended users of the system.

Speech material, talker and listener effects are perhaps the most apparent of the types of effects that must be controlled in any attempt to assess speech quality. They are by no means all of the effects about which one must be concerned, however. Busch and Eldredge (1972) have shown, for example, that results can be significantly affected by such incidentals as the time intervening between the presentation of successive test items and the way in which the subject makes his response. Moreover, one apparently cannot safely assume that the effects of such variables will combine additively with those of other variables of greater interest, inasmuch as interactions are sometimes found. McGee (1964, 1965) has reported order effects in two different experiments involving pair comparison tasks, the second member of the pair being favored in each case. Rothauser, Urbanek and Pachl (1968) also seem to recognize such order effects.

8. THE POSSIBILITY OF OBJECTIVE MEASURES OF SPEECH QUALITY

The collection of subjective evaluation data is costly and time-consuming. A more efficient way of determining the quality of speech would be to infer it from objective measurements made on the speech signal itself. The problem is that not enough is known about the relationship between objectively measurable properties of speech and its perceived quality to assure the effectiveness of this approach. What is needed is a model that relates objective and perceptual properties of speech in an unambiguous way.

Such a model would be an invaluable aid, both to developers of speech processing and communication systems and to teachers of individuals with various types of speech impairments. It is difficult to see how, without such a model, efforts to improve the performance of speech-production systems (whether artificial or human) can be given effective guidance. Consider, for example, the problem faced by the teacher of a deaf child. The teacher may know that the child's speech is grossly defective, perhaps both in quality and intelligibility. Furthermore, he may be able to identify some fairly specific deficiencies. He may know that the pitch is generally too high and monotone, that the child speaks too slowly and without adequate temporal differentiation between syllables that should receive primary stress and those that should be unstressed, that certain phonemes are omitted or misarticulated, and so on. But even with this type of knowledge of what is wrong, it is not clear how one should go about trying either to increase the intelligibility of the speech, or to improve its overall quality. It is quite certainly not the case that where one starts does not matter. It seems highly likely that some deficiencies are more detrimental to intelligibility or quality than are others; but little is known concerning specifics in this regard.

Ideally one would like to be able to infer speech quality and other perceptual characteristics of an utterance from a set of objective measurements. But the things that can vary in a speech sample and the sorts of measurements that can be made are discouragingly numerous. Often there is disagreement among investigators regarding how what appear to be simple measurements should be made: even such a seemingly straightforward characteristic of speech as its level is still measured in a variety of ways, and a single standard measurement technique has yet to be agreed upon (Brady, 1971). Moreover, the best that one can hope to do with a model that predicts quality from objective properties of speech is bounded above by the degree to which quality--as represented by listener assessments--is in fact determined by stimulus properties, as opposed to listener variables. Such problems notwithstanding, it seems reasonable to attempt to develop such predictive models and the time appears to be right for doing so. The involvement of computers in speech compression and speech synthesis procedures should facilitate the development of such models, because it provides the opportunity to obtain large numbers of measures on the speech signal and to subject them to many different types of analyses.

In the case of vocoded, or synthesized, speech an alternative to making measurements on the speech signal is that of attempting to predict quality from the parameters of the vocoder or synthesizer, or, in the case of linear predictive coding, on some measures of difference between the preprocessed and the vocoded speech. Some work along these lines has been done. In particlar, three recent studies have been reported, and the results look promising.

All three studies share the same approach: they attempt to measure the spectral error introduced by LPC vocoding (the method can be extended to other types), on a frame-by-frame basis, and then pool the error across all the frames in an utterance to arrive at a single number representing the vocoder's ability to represent the speech spectrum accurately. This objective measure is then correlated with subjective estimates of quality. Meister and Wiggins (1976) developed a measure that involved (a) finding the difference between the log-area ratios calculated from the reflection coefficients at the input to the re-synthesizer and those calculated from the reflection coefficients obtained by re-analysis of the synthesizer's output, on a frame-by-frame basis; (b) weighting these differences by the average frame power (since errors in loud speech should outweigh those in soft speech); (c) taking the mean error across all frames, and adding to it the mean of the 20 largest terms (since large errors should be more important than small ones); and finally (d) taking the difference between the two figures so arrived at, for the two systems being compared, to yield their Quality Comparison Measure. They then tested their method with a set of twelve pairs of vocoders. Unfortunately, they failed to specify what their twelve pairs were, but one pair differed only in whether a Hamming or a rectangualr window was used for sampling the waveform, and a second differed only in the analysis method used. Coding issues were not addressed. Their measure (which was developed by post-hoc analysis) gave highly significant correlations with subjective results.

Makhoul, Viswanathan and Russell (1976) argue that most of the significant degradation of speech quality in narrow-band LPC vocoders occurs during encoding, rather than during analysis and resynthesis, since heavy quantization of the filter coefficients is necessary to achieve the desired low bit rate. They therefore compared the spectra represented by the encoder, with those used by the synthesizer after interpolation. The test is thus "inside" the vocoder. A second requirement considered vital by Makhoul et al is that the distance measures used to compare the two spectra (and for many other purposes) should relate to known perceptual constraints. They tried a variety of frequency weightings, including spectral intensity, frequency derivative, articulation index, and perceived loudness weightings, but found that none of the measures accurately predicted subjective preferences under all conditions.

In a second paper Viswanathan, Makhoul and Russell (1976) point out that the traditional spectral distance measure, based on mean squared error, treats spectral errors symmetrically -- that is, an error in one direction is equivalent to an equal error in the other direction. This conflicts with perceptual results, which have shown that an error is much more noticeable if it reduces the separation of two adjacent formants, than if it increases the separation -- that is, errors should not be treated symmetrically. They describe a distance measure that has the required property, and work is under way to develop and test it further.

9. QUALITY AND INTELLIGIBILITY

We have been concerned so far with the problem of assessing the quality of speech independently of its intelligibility. The rationale for this restriction of our attention is based on two assumptions: (a) that speech-processing systems of the sort that investigators are often interested in evaluating have progressed to the point that intelligibility is not a major issue, and (b) that even speech that is highly intelligible may differ qualitively in ways that have implications for the acceptability of speech-processing systems to their users. As was pointed out in the introduction, however, the distinction between intelligibility and speech quality is not a sharp one. One might argue that since quality tests are needed only to distinguish between systems with equal (and usually very high) intelligibility, they can be regarded as simply expanding the top end of the intelligibility scale. As we have seen, however, quality tests have the disadvantage that they require subjective judgements rather than responses that can be objectively scored as right or wrong. An alternative approach is to expand the upper end of an objective intelligibility test, by making the test more difficult. We now turn to a consideration of two ways in which intelligibility testing may be important, even for speech-processing systems whose output is assumed to be highly intelligible.

One reason for such testing is the fact that speech that is equally (and highly) intelligible under favorable listening conditions is not necessarily equally resistant to various forms of degradation. This is the general problem of ceiling effects in performance testing. Engineering psychologists have long recognized the fallacy in assuming that because two systems operate equally well under close-to-ideal conditions they will continue to operate equally well under adverse conditions. In keeping with this reasoning, the testing of communications systems has often included attempts to determine how well a system performs under various conditions that would be expected to affect it detrimentally. Typically, the factors that are manipulated in these tests are variables that affect the signal in some direct way, e.g., the attenuation of signal strength, or the addition of masking noise to the circuit (Becker and Kryter 1975).

Nakatani and Dukes (1973) have argued that if there is any perceivable difference in quality between two systems that can lead to a subjective preference for one system over the other, that superiority in quality should be translatable into an intelligibility advantage under some set of conditions. They proposed two sets of such conditions, of which only one predicted subjective quality ratings successfully. Their "Q-measure" is obtained by comparing the Signal/Interference level yielding 50% intelligibility for the degraded speech with the S/I level yielding 50% intelligibility for high-quality reference speech. The interference in both cases was an irrelevant message, processed through the <u>same</u> system under test. When both signal and interference speech were presented to subjects binaurally, (the 2-Channel condition) the Q-measure was found to correlate highly with subjective ratings of the systems under test. However, they also found that the Q-measure was not an adequate predictor of quality, when the target was presented binaurally (i.e. yielding a central fused image), and two different Interference sentences were presented simultaneously, one to each ear (the 3-Channel condition). Unfortunately, the 2-channel test was run on a smaller set of systems than the 3-channel test, and some of the excluded systems were those that caused the poor correlation in the 3-channel test. On the other hand, the 2-Channel test yielded Q-measures with considerably less dispersion than the 3-Channel test.

Another possible method for increasing the difficulty of the intelligibility testing task for the listener, is to reduce the contextual information that he has available to help interpret the speech, or by imposing other tasks on him that must be performed simultaneously, thus presumably diverting attention from the speech-perception task. The intelligibility of the speech produced by different systems should be considered equivalent only if it decreases at the same rate for the two systems, as listening conditions are made progressively worse. Consideration of this factor is particularly important, of course, in the case of systems that are likely to be used in operational situations that are less than ideal.

A further reason for the use of intelligibility testing on "highly intelligible" speech is the fact that such testing may provide some useful information concerning the capability of a system to represent specific speech sounds. The evidence that listeners normally make use of context to disambiguate some aspects of even a "good" speech signal is very compelling. Listener identification of vowel sounds in the context /hVd/, even when they have been very carefully recorded on high-fidelity equipment, tends to be something less than 100% (Peterson & Barney, 1952). Given that context is used pervasively in understanding running speech, the fact that a listener can correctly transcribe an utterance does not guarantee that the speech sounds comprising the utterance are recognizable individually. Or, the fact that two systems produce connected speech that is equally intelligible does not guarantee that those two systems are equivalent in terms of their ability to produce specific sounds. There is, in short, a reason for doing phoneme-specific intelligibility testing on speech, the overall intelligibility of which is high. Particularly is this true when there is some a priori basis to suspect that the systems of interest may differ in their abilities to produce specific sounds (Stevens, 1962).

9.1 Quantification of Intelligibility

The problem of quantifying intelligibility has received a fair amount of attention from speech scientists. We make no effort to review here the work that has been done on this problem, except to cite a few studies that make the point that any measure of intelligibility is interpretable only with reference to the procedure by which it was obtained. It is not enough to say that a speech sample is intelligible, or unintelligible; one wants to know how intelligible (or unintelligible) it is, and to whom, and under what conditions.

Degree of intelligibility typically is reported in percentage points. The percentage usually indicates the number of words that are correctly recognized (perhaps with a correction for guessing) relative to the total number comprising a test. Sometimes all the words of an utterance constitute test words; sometimes only one or a few of them do, while the other words comprise a "carrier" and provide a context for the test word(s). Sometimes the listener is provided with a set of alternative possibilities from which to select the test word(s); sometimes, he is expected to make the identification without such help.

There is extensive evidence that the resulting intelligibility score that one obtains may depend very much on such details. Words are more easily identified in noise when the listener is provided with a set of alternatives from which to select his response than when he is not. A decision between only two alternatives is possible at a signal to noise ratio of -14 db whereas a selection among English monosyllables requires a signal-to-noise ratio of +4 dB for the same score (Miller, Heise, and Lichten, 1951). Words presented in a meaningful linguistic context are reported with greater accuracy than are words presented in isolation (Hirsh, Reynolds & Joseph, 1954; Miller, Heise & Lichten, 1951), the amount of facilitation depending on the degree to which test items are predictable from the context (Stowe, Harris, and Hampton, 1963; Kalikow, Stevens & Elliot 1976).

Although percentage of words identified correctly is the most common measure of intelligibility in use, it is a relatively gross measure. It tells one nothing, for example, about the degree of difficulty that a listener may have experienced in interpreting the speech signal, or of his confidence that he has, in fact, interpreted it correctly. Hecker, Stevens and Williams (1966) proposed that other measures should perhaps be developed that could reflect this type of difference, and have performed one preliminary test on the usefulness of reaction time as such a measure. They found a monotonic relationship between reaction time and percent words correct, as did Pollack and Rubenstein (1963) in an earlier study: as the signal-to-noise ratio was decreased, percent correct decreased and reaction time increased. Although reaction time was slightly less for correct than for incorrect responses, the fact that it increased with decreasing signal-to-noise ratio in both cases led Hecker et al to conclude that the percent-correct measure of intelligibility and reaction time are independent to some degree.

Another possible approach to the measurement of intelligibility, and perhaps of speech quality as well, is that of assessing the effectiveness of the speech in communication situations (Richards and Swaffield, 1958). Chapanis (1973; 1975) and his colleagues have recently used problem-solving tasks on which two people must cooperate, as a vehicle for studying the effectiveness of various means of communication between the collaborators. The results of his experiments have demonstrated the utility of speech as opposed to non-speech methods of communication. Becker (1975) has proposed the use of a similar method for assessing the communicative utility of processed speech. Percentage of words identified correctly would not be an appropriate performance measure in this case, of course; rather one would use such measures as the amount of time required to solve a problem, the number of words or utterances that were spoken, the number of requests for repetition of some part of an utterance, and so on. Hiller (1976) has also reported a new method for measuring utility of communication. He measures the time taken for a text to be transmitted exactly through a channel. Every error necessitates a repetition, and increases the time.

It is also possible, of course, to take the real-life use of a system as a testing situation. In several studies of the effects of transmission delays on telephone conversations (Klemmer, 1967), participating subjects had delays switched into their office telephones whenever they made calls within the company. If communication was too difficult, one could dial a 3 to remove the delay. The distribution of conversation durations before the escape was requested provided a measure of acceptability.

10. CONCLUDING COMMENT

Speech production and speech perception are extremely complex processes and neither is yet thoroughly understood. It perhaps should not be surprising, therefore, that the assessment of speech quality has proven to be a difficult task. The difficulty stems in part from the subjective and somewhat inscrutable nature of human preferences, in part from the fact that speech--even highly intelligible speech--can vary qualitatively in so many ways, and in part from the fact that this variability is determined by numerous factors. Speech remains a preferred way of communicating among people, however, and the advent of computer-mediated communication systems with the attendant proliferation of potential uses of processed speech increases the importance of finding more efficient methods of quality assessment. To the extent that the search for such methods is successful, it should also have a beneficial inpact on the task of evaluating the quality of unprocessed speech, and thereby facilitate the remediation of speech-production disabilities.

Bayless, J. W., Campanella, S. J., & Goldberg, A. J. Voice signals: bit-by-bit IEEE Spectrum, October 1973, 28-34.

Beasley, D. S., Zemlin, W. R. & Silverman, F.H. (1972) Listeners' judgements of sex, intelligibility, and preference for frequency-shifted speech. Percept. Motor Skills 34(3), p782.

Becker, R.W. & Kryter, K. D. Assessment of acceptability of digital speech communication systems. Annual Technical Report, Stanford Research Institute, Project 3843, May 1975. Sponsored by Defense Advanced Research Projects Agency.

Boothroyd, A., Nickerson, R.S. & Stevens, K. N. Temporal patterns in the speech of the deaf -- An experiment in remedial training. Clark School for the Deaf, Research Department Report No. S.A.R.P. #15, 1974.

Brady, P.T. Need for standardization in the measurement of speech level. Journal of the Acoustical Society of America, 1971, 50, 712-714.

Bricker, P.D. Study of the nature of speech-transmission-circuit quality by means of the direct production of equivalent impairments. Journal of the Acoustical Society of America, 1963, 35, 1899.

Busch, A. C. & Eldredge, E. Effects of stimulus time interval, response mode and test material for intelligibility testing. Proceedings, Conference on Speech Communication and Processing, Institute of Electrical and Electronics Engineers and Air Force Cambridge Research Laboratories, April 1972, 183-186.

Carroll, J.D. Individual differences and multidimensional scaling. In R.N. Shepard, A.K. Romney, & S. Nerlove (Eds.), Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. Vol. 1 Theory. New York: Seminar Press, 1972, 105-155.

Cederlof, R., Jonsson, E., & Sorensen, S. On the influence of attitudes to the source on annoyance reactions to noise: a field experiment. Nordisk Hygienisk Tidskrift. 1967, 48, 46-49.

Chapanis, A. The communication of factual information through various channels. Information Storage and Retrieval, 1973, 9, 215-231.

Chapanis, A. Interactive human communication. Scientific American, 1975, 232, 36-42.

Coolidge, O.H., & Reir, G.C. An appraisal of received telephone speech volume. Bell System Technical Journal, 1959, 38, 877.

Flanagan, J.L. Speech Analysis, Synthesis, and Perception. New York: Springer-Verlag, 1972.

Gabrielsson, A. & Sjogren, H. Adjective ratings and dimension analyses of perceived sound quality of hearing aids I. Report #75, Technical Audiology, Karolinska Institutet, Stockholm, 1974.

Gabrielsson, A. & Sjogren, H. Adjective ratings and dimension analyses of perceived sound quality of hearing aids II. Report #75, Technical Audiology, Karolinska Institutet, Stockholm, 1975a.

Gabrielsson, A. & Sjogren, H. Similarity ratings and dimension analyses of perceived sound quality of hearing aids. Report #76, Technical Audiology, Karolinska Institutet, Stockholm, 1975b.

Gabrielsson, A., Rosenberg, U., & Sjogren, H. Judgments and dimension analyses of perceived sound quality of sound reproducing systems. Journal of the Acoustical Society of America, 1974, 55, 853-861.

Green, P.E., & Rao, V.R. Multidimensional Scaling - An In-Depth Comparison of Approaches and Algorithms. New York: Holt, Rinehart, and Winston, 1971.

Grether, C.G., & Stroh, R.W. Subjective evaluation of differential pulse-code modulation using the speech "goodness" rating scale," IEEE Transactions on Audio and Electroacoustics, 1973, AU-21, 179-184.

Hecker, M.H.L., Stevens, K.N., & Williams, C.E. Measurements of reaction time in intelligibility tests. Journal of the Acoustical Society of America, 1966, 39, 1188-1189.

Hecker, M.H.L., & Williams, C.E. Choice of reference conditions for speech preference tests. Journal of the Acoustical Society of America, 1966, 39, 946-952.

Heyduk, R.G. Rated preference for musical compositions as it relates to complexity and exposure frequency. Perception & Psychophysics, 1975, 17, 84-91.

Horii, Y., House, A.S., & Hughes, G.W. Making noise with speech envelope characteristics for studying intelligibility. Journal of the Acoustical Society of America, 1971, 49, 1849.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 1965, 37, 158-166.

Hudgins, C. V. Speech intelligibility tests: A practical program. Volta Review, 1943, 45, 5-6, 52,54.

Hudgins, C. V. A method of appraising the speech of the deaf. Volta Review, 1949, 51, 597-601, 638.

Huggins, A.W.F. & Nickerson, R.S. Some effects of speech materials on vocoder quality evaluations. Journal of the Acoustical Society of America, 1975, 58, S129, A.

IEEE Recommended Practice for Speech Quality Measurements, IEEE Standards 297, April 1969, Also IEEE Trans. Audio and Electroacoustics AU-17, 225-246, 1969.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. BBN Report No. 3370, 15 September 1976, Submitted for publication, Journal of the Acoustical Society of America.

Kerrick, J.S., Nagel, D.C., & Bennett, R.L. Multiple ratings of sound stimuli. The Journal of the Acoustical Society of America, 1969, 45, 1014-1017.

Klemmer, E.T. Subjective evaluation of transmission delay in telephone conversations. Bell System Technical Journal, 1967, 46, 1141-1147.

Kramer, E. Judgement of personal characteristics and emotions from non-verbal properties of speech. Psychological Bulletin, 1963, 60, 408-420.

Makhoul, J., Viswanathan, R., & Russell, W. A framework for the objective evaluation of vocoder speech quality. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, April, 1976, 103-106.

Makhoul, J.I., & Wolf, J.J. Linear prediction and the spectral analysis of speech. BBN Report # 2304, 1972.

Martony, J. & Franzen, O. Formant transitions and phoneme lengths as constituents of speech naturalness. Quarterly Progress and Status Report 1-66, 1966. Speech Transmission Laboratory, Royal Institute of Technology, Stockholm.

McGee, V.E. Semantic components of the quality of processed speech. Journal of Speech Hearing Res., 1964, 7, 310-323.

McGee, V.E. Determining perceptual spaces for the quality of filtered speech. Journal of Speech Hearing Res, 1965, 8, 23-28.

McDermott, B.J. Multidimensional analysis of circuit quality judgments. *Journal of the Acoustical Society of America*, 1969, *45*, 774.

Meister, S., & Wiggins, R.H. Quality comparison measure for linear predictive systems. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, April 1976, p. 107-109.

Miller, G.A., Heise, G.A., & Lichten, W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 1951, *41*, 329.

Miller, J.M. Directions for basic research in the development of protheses for the hearing impaired. Paper presented at AAAS meeting, Boston, February 18-24, 1976.

Mostofsky, D.I. Alternative strategies in the evaluation of speech systems. AFCRL, Bedford, Mass., Report AFCRL - 69 - 0357, August 1969.

Munson, W.A., & Karlin, J.E. Isopreference method for evaluating speech-transmission circuits. *Journal of the Acoustical Society of America*, 1962, *34*, 762-774.

Nakatani, L.H., & Dukes, K.D. A sensitive test of speech communication quality. *Journal of the Acoustical Society of America*, 1973, *53*, 1083-1092.

Nye, P. W., Ingemann, F., & Donald, L. Synthetic speech comprehension: A comparison of listener performances with, and preferences among, different speech forms. Status Report on Speech Research, SR-41, 117-126, 1975, Haskins Laboratories, New Haven, Conn.

Osgood, C.E. The nature and measurement of meaning. *Psychology Bulletin*, 1952, *49*, 197-237.

Pachl, W.P., Urbanek, G.E., & Rothauser, E.H. Preference evaluation of a large set of vocoded speech signals. *IEEE Transactions of Audio-Electoacoustics*, 1971, *AU-19*, 216-224.

Peterson, G.E., & Barney, H.L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175-184.

Pollack, I., & Rubinstein, H. Response times to known message sets in noise, language and Speech. 1963, *6*, 57-62.

Richards, D.L., & Swaffield, J. Assessment of speech communications links. *Prococeedings Institute Electrical Engineering B*, 1958, *106*, 77-92.

Robinson, D.W., Bowsher, J.M., & Copeland, W.C. On judging the noise from aircraft in flight. Noise, 1963, 186-203.

Rothauser, E.H., & Urbanek, G.E. New reference signal for speech-quality measurements. Journal of the Acoustical Society of America, 1965, 38, 940 (A).

Rothauser, E.H., Urbanek, G.E., & Pachl, W.P. Isopreference method for speech evaluation. Journal of the Acoustical Society of America, 1968, 44, 408-418.

Rothauser, E.H., Urbanek, G.E., & Pachl, W.P. Comparison of preference measurement methods. Journal of the Acoustical Society of America, 1971, 49, 1297-1308.

Schroeder, M.R. Vocoders: Analysis and synthesis of speech. Proceedings of the IEEE, 1966, 54, 720-734.

Schroeder, M.R. Reference signal for signal quality studies. Journal of the Acoustical Society of America, 1968, 44, 1735.

Stevens, K.N. Simplified nonsense - syllable tests for analytic evaluation of speech transmission systems. Journal of the Acoustical Society of America, 1962, 34, 729 (A).

Stevens, K.N., Nickerson, R.S., Rollins, A., & Boothroyd, A. Use of a visual display nasalization to facilitate training of velar control for deaf speakers. BBN Report # 2899, 1974.

Stowe, A.N., Harris, W.P., & Hampton, D.B. Signal and context components of word-recognition behavior. Journal of the Acoustical Society of America, 1963, 35, 639 - 644.

Stratton, W. D. Intonation feedback for the deaf through the tactile sense. M.I.T. Thesis, 1973.

Tedford, W.H., Jr., & Frazier, T.V. Further study of the isopreference method of circuit evaluation. Journal of the Acoustical Society of America, 1966, 39, 645-649.

Viswanathan, R., Makhoul, J., Russell, W. Towards perceptually consistent measures of spectral distance. International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, April 1976, 485-488.

Voiers, W.D., Sharpley, A.D., & Hehmsoth, C.J. Research on diagnostic evaluation of speech intelligibility. Air Force Cambridge Research Labs Report #AFCRL-72-0694, 1972.

Zemlin, W.R., Daniloff, R.G., & Skinner, T.H. (1968) The Difficulty of Listening to Time-Compressed Speech J. Sp. Hearing Res. 11, p875-881